Chapter 6. Conclusions and future prospects

This dissertation described four distinct projects that share the common theme of unravelling the genetic basis of complex diseases. In Chapter 1, I gave a historical perspective of our understanding of the genetics of complex traits, from early 20th century efforts at reconciling Mendel's laws with the inheritance of quantitative phenotypes, to attempts throughout the 1980s to early 2000s at identifying complex disease risk loci via linkage scans, and finally to the success of GWAS from the mid-2000s up to the present day. In Chapters 2, 3 and 4, I described such locus discovery projects in PBC, PSC and IBD respectively, much of which was undertaken using the Immunochip custom genotyping array. The dense SNP content of the array has allowed for greater refinement across risk loci, while its low cost has enabled powerful locus discovery projects and cross-phenotype comparisons in very large sample sizes. Once a set of risk loci for a particular disease is found, there is also the question of what to do next. In Chapter 5, I described a simple method of combing disease risk loci with tissue-relevant functional genomic datasets in order to identify candidate genes at these risk loci, as well as potential mechanisms through which they mediate disease.

Pick up any issue of a reputable genetics journal from the past seven years and it may seem that locus discovery in complex traits is routine, if a little tedious for some. Visiting the NHGRI GWAS Catalog (Hindorff *et al.*, 2014) leaves one in no doubt, with 1,961 publications listed and 14,012 reported associated variants as of September, 2014. In the following pages, I will discuss the general lessons learnt from these types of studies, and then will look to future prospects and challenges for locus discovery, understanding biology, and ultimately translating these findings into better treatment outcomes.

6.1 Effect sizes, power and the genetic architecture of complex traits

For genetic studies of complex traits, sample size is key. With few exceptions (e.g. HLA region in immune-mediated disorders and *NOD2* in CD), the effect of individual common genetic variants on disease risk is modest – allelic odds ratios are typically less than 1.2, and almost always less than 1.5. Robustly identifying these loci requires a combination of large sample sizes, genome-wide coverage and strict statistical criteria for determining significance – three aspects that were overlooked in early linkage and candidate genes studies. Figure 6.1 illustrates the appreciation of the need for large samples in order to robustly identify susceptibility loci – there is clear positive correlation between the number of loci discovered and the sample size of the study.

Chapters 2, 3 and 4 described the largest genetic studies to date for PBC, PSC and IBD respectively in terms of the number of samples recruited. However, the total proportion of variation in disease liability explained by these loci is still modest. Figure 6.2 illustrates the relationship between the cumulative proportion of variance explained and the strength of association. Several conclusions may be drawn from this graph. Firstly, extrapolating these curves clearly suggests that many more risk loci will be discovered as sample sizes get larger, with the total number increasing at an exponential rate (with respect to sample size) while the effect size of each individual locus will get ever smaller (Park *et al.,* 2010). These effect size distributions suggest that there are potentially thousands of susceptibility loci underlying these complex disorders.

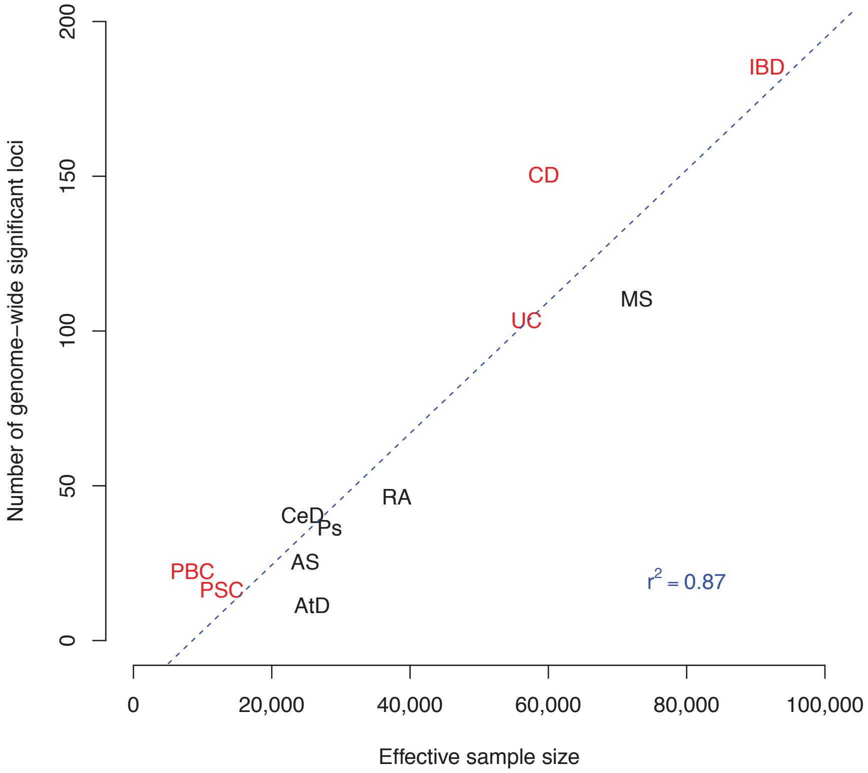

Figure 6.1. Effective sample size vs. number of genome-wide significant risk loci across GWAS and Immunochip studies of nine immune-mediated disorders. Diseases denoted in red formed Chapters 2-4 of this dissertation. The dashed blue line indicates the best fitting line estimated from least squares regression. The effective sample size denotes the cohort with an equal number of cases and controls that have an equivalent power as the sample sizes reported in the original study. This was estimated by iterating sample sizes with a 1:1 case:control ratio until it arrives at the same non-centrality parameter in power calculations as the reported sample size (Purcell et al., 2003). The studies listed are – AS: ankylosing spondylitis (International Genetics of Ankylosing Spondylitis Consortium, 2013), AtD: atopic dermatitis (Ellinghaus et al., 2013a), CeD: coeliac disease (Trynka et al., 2011b), CD: Crohn's disease, UC: ulcerative colitis, IBD: inflammatory bowel disease (Chapter 4), MS: multiple sclerosis (International Multiple Sclerosis Genetics, 2013), PBC: primary biliary cirrhosis (Liu et al., 2012), Ps: psoriasis (Tsoi et al., 2012), PSC: primary sclerosing cholangitis (Liu et al., 2013), RA: rheumatoid arthritis (Eyre et al., 2012).

Secondly, the decreasing effect sizes also raises questions about how much of total heritability can be explained by common variants. Assuming that narrow-sense heritability in Crohn's disease is 50% (Ahmad *et al.,* 2001), extrapolating the risk loci to 20,000 independent common variant associations will still explain less than half of this heritability (Franke *et al.,* 2010). A similar estimate was arrived at in Lee *et al.* (2011), where the variance explained in Crohn's disease risk tagged by all genotyped variants was only 22%. This suggests that untyped variants (especially rare variants poorly tagged on genotyping arrays) will contribute to the remaining heritability, though that heritability estimates were overestimated in the first place cannot be ruled out. Due to the disease being rare, accurate disease prevalence is hard to estimate. Similarly, familial recurrent risk estimates are often based on ascertained families will multiple affected individuals, potentially overestimating the true familial risk in the population.

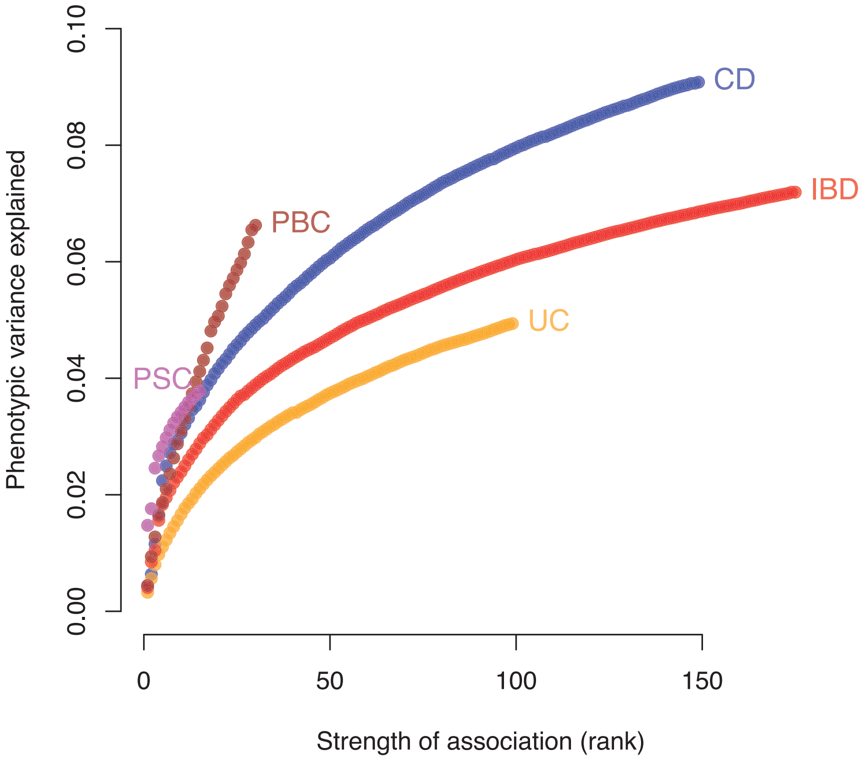

Figure 6.2. Cumulative proportion of variance in disease liability explained by the genome-wide significant loci identified in Chapters 2-4. For PBC and PSC, SNPs on the x-axis are ranked by and plotted by association P-value. For CD, UC and the two combined (IBD), SNPs are ranked by the decreasing MANTRA log10 Bayes factor association signal.

The different trajectories for each of the diseases in Figure 6.2 reflects the different underlying genetic architectures for these disorders and their tractability to the GWAS approach. For instance, the variance explained by the 26 loci associated with PBC is more than double the equivalent number in UC. While some of these differences may be due to winner's curse (the sample size for PBC was much smaller and there has yet to be any follow-up studies), this also raises interesting questions about factors that shape these differences.

Before discussing these factors, it is interesting to compare the distribution of effect sizes of variants associated with immune-mediated disorders with those from other complex traits. In general, genetic studies for immune-mediated disorders have offered much greater bang-for-genotyping-buck in terms of the number of risk loci discovered and variance explained than other classes of disorders. For instance, the PBC study described in Chapter 2 identified 22 genome-wide significant loci with a sample size of ~2,800 cases and 8,500 controls. In contrast, it required over 5,500 cases and 9,000 controls to identify a single variant associated with endometriosis (Painter *et al.,* 2011). For psychiatric disorders, the story is just as sobering. Despite heritability estimates of ~30-40% in major depressive disorder, only a single borderline genome-wide significant signal was identified in a meta-analysis that included over 16,000 cases and 60,000 controls (Major Depressive Disorder Working Group of the Psychiatric, 2013). Nevertheless, even for these classes of disorders, risk loci will eventually be identified given large enough sample sizes. In schizophrenia, it required 8,000 cases and 19,000 controls to implicate a single locus in disease risk (Shi *et al.,* 2009). Five years later, a GWAS meta-analysis that included 36,000 cases and 113,000 controls increased the number of risk loci to 108 (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014).

Why then, is locus discovery in immune-mediated diseases such as PBC more tractable to the GWAS approach than diseases such as schizophrenia and endometriosis? One explanation is related to the hygiene hypothesis – that evolutionary adaptations have not caught up to the rapidly changing environment. Natural selection leaves its indelible footprints on the frequencies of functionally relevant alleles. For much of human history, it's likely that exposure to a range of pathogens was the norm, and alleles that best defend the host from infection were selected for and rose in frequency throughout the population. In the modern world, vaccinations, better nutrition and awareness of hygiene have greatly reduced our exposure to antigens, leading to an imbalance in the immune system that favours chronic inflammatory conditions (Sironi and Clerici, 2010). It is hypothesised that those same alleles that once protected us from infection are now also those that make us most susceptible to autoimmune disorders.

A second explanation of differences in GWAS tractability is the amount of phenotypic and genetic heterogeneity that underlies complex traits. The presence of heterogeneity in genetic association studies reduces power to detect association and underestimates the effect sizes of risk variants. At the biological level, what is classified as a single disorder may be a result of combinations of different molecular processes, each with its own set of genetic and environmental levers, yet all with similar phenotypic presentations. This may especially be true for psychiatric disorders, where a yes/no diagnosis is still often based on whether a patient shows any *x* number of descriptive symptoms out of a list of *y* (Angst, 2007), resulting in potential for misclassification of cases and controls. This is largely because the most useful biological categories or dimensional categories are still unknown, and a better understanding of the genetic basis of these disorders will help give a clearer picture of disease pathogenesis and diagnoses. Contrast this to an autoimmune disorder such as PBC, where diagnosis is largely based on blood tests and the presence of a specific set of antibodies.

These two hypotheses are not mutually exclusive, and indeed are both likely to play a role in shaping the genetic architecture of complex traits. Future efforts at unravelling this genetic architecture will involve a combination of array-based and sequencing approaches in ever-larger sample sizes. For the remainder of this chapter, I will discuss these approaches, their potential challenges, and ultimately, prospects for translating what we've learnt from locus discovery into more effective treatment outcomes.

6.2 Future prospects for complex disease genetics

6.2.1 Array-based approaches

Genome-wide association studies predominantly focused on identifying common variant associations (variants with minor allele frequencies greater than 5%). There are good economic reasons for why this was the case. There are only so many SNPs that can fit on a genotyping chip, and given the patterns of linkage disequilibrium in the population, the majority of the ~5 million common variants in the genome can be tagged by a selection of ~500,000 SNPs (Barrett and Cardon, 2006; International HapMap *et al.,* 2007). Array-based studies with ever-larger sample sizes will continue to play a role in locus discovery. This is perhaps best exemplified by the UK Biobank's ongoing efforts to genotype their ~450,000 samples on a custom genome-wide genotyping microarray with ~800,000 variants. Individuals recruited to the UK Biobank underwent a range of diagnostic measures and will have their health tracked throughout their lifetime, providing an invaluable resource in the study of complex disease.

The design of the Immunochip, along with similar arrays such as the Metabochip (for metabolic and cardiovascular risk loci) (Voight *et al.,* 2012) and iCOGS (for various cancers) (Sakoda *et al.,* 2013), was also primarily motivated by economics. The ability to include thousands of SNPs for deep replication, high-density regions for fine-mapping, and the genotyping of over 150,000 individuals across multiple disease cohorts (and the sharing of population controls) meant that the Immunochip, at ~$40/sample, was a much more cost-effective platform for locus discovery and fine-mapping than alternative technologies at the time (e.g. Sequenom plexes, whole-genome arrays, pull-down sequencing) (Jostins, 2012).

There are, of course, several limitations to custom high-density arrays such as the Immunochip. Obvious pitfalls include the lack of coverage genome-wide and the ascertainment of variants only present in European populations. Additionally, while ~240,000 variants were initially selected for inclusion on the Immunochip, 196,524 made it onto the final array. Running a typical quality control protocol will reduce this even further to 130,000-140,000 variants (Liu *et al.,* 2012; Liu *et al.,* 2013), resulting in a total array design success of ~60%. Technical failures explain the majority of these exclusions. SNPs in high-density regions were selected from those identified in the 1000 Genomes Pilot dataset using low-coverage sequencing, such that many of these variants (in particular rare variants) are poorly characterised, either due to being falsely called in the first place and/or poor probe design. Moreover, many variants were missed all together. As demonstrated in Chapters 2 and 3, imputation using the subsequently much larger 1000 Genomes Phase I reference panel almost doubled the number of variants in the high-density regions. Nevertheless, chip design continues to improve, and there are now several custom chips currently being developed or in the analysis phase – e.g. the Exome Chip for coding variants, the "African Power Chip" for African-specific variants, and the "Psych Chip" for risk variants identified in psychiatric disorders. In addition, current genome-wide arrays such as the Illumina Omni2.5 and Omni5 are supplemented with 200,000 and 500,000 custom variants respectively to fit with each researcher's requirements.

6.2.2 Sequencing approaches for rare variant studies

How then, given the state of technology and what we understand about the genetic architecture of complex traits, should one design a locus discovery experiment today? Array-based technologies (whole-genome and targeted arrays) are likely to remain the most cost-effective and efficient methods for identifying common variant associations, though a complete survey of genetic variation in an individual will require high coverage (greater than 30X) whole-genome sequencing – currently costing 1-2 orders of magnitude more per sample than genotyping arrays. These sequencing approaches will be able to capture rare variants (those with minor allele frequencies less than 1%), which are poorly captured on arrays. While most genetic variation in an individual is at common sites, the total number rare variants in the population far outnumber common variants (Keinan and Clark, 2012). In chapter 1 section 1.4.3, I outlined theoretical reasons why rare variants are likely to play a role in complex disease, and highlighted recent sequencing studies in known risk loci to identify rare variant associations (Hunt *et al.,* 2013; Rivas *et al.,* 2011).

These targeted sequencing studies identified very few novel independent rare variant signals (and that the common variant associations are not driven by nearby rare variants), highlighting the need for larger sample sizes for these types of studies. Under certain assumptions about the effect size distribution of rare variants and selection pressures, well-powered studies may require cohorts of more than 25,000 cases and an equal number of controls, along with equally large numbers for replication (Zuk *et al.,* 2014). Moreover, given the importance of non-coding variation in complex disease risk, there is also a need for whole-genome approaches.

While high-coverage sequencing is still prohibitively expensive, there is currently great potential for low-coverage whole-genome sequencing approaches (less than 6X) as a powerful and cost-effective alternative. In low-coverage sequencing, rare variants are discovered and jointly called across many thousands of individuals, and LD-based imputation methods are used to refine genotype calls. For instance, for a SNP with frequency 0.2% to be discovered, over 2000 individuals need to be sequenced at 30X coverage (60,000 genomes). In contrast, the same SNP can be identified in ~3000 individuals sequenced at 4X (12,000 genomes) – a five fold reduction in sequencing cost (Li *et al.,* 2011). With more sequenced individuals comes greater power to detect associations. Large cohorts of low-coverage sequenced individuals can also be used as reference panels to impute rare variants into new and existing GWAS datasets at much greater accuracy than existing panels. Over the course of 2014-15, it is expected that over 30,000 individuals will be sequenced at low-coverage (www.haplotype-reference-consortium.org). Imputing the millions of new variants discovered from this set into ~25,000 IBD cases (of which ~15,000 have already been genotyped as part of GWAS) will, for the first time, enable dection of association to SNPs with frequencies in the order of 0.1-1% and ORs of 2-3 (Figure 6.3). This sequencing plus imputation approach was demonstrated in a recent study in type 2 diabetes, were variants discovered by sequencing 2,630 samples were imputed in 11,114 cases and 267,140 controls (Steinthorsdottir *et al.,* 2014). The study identified risk variants at several variants with frequencies between 0.65% and 1.5% and ORs of 1.5-2.

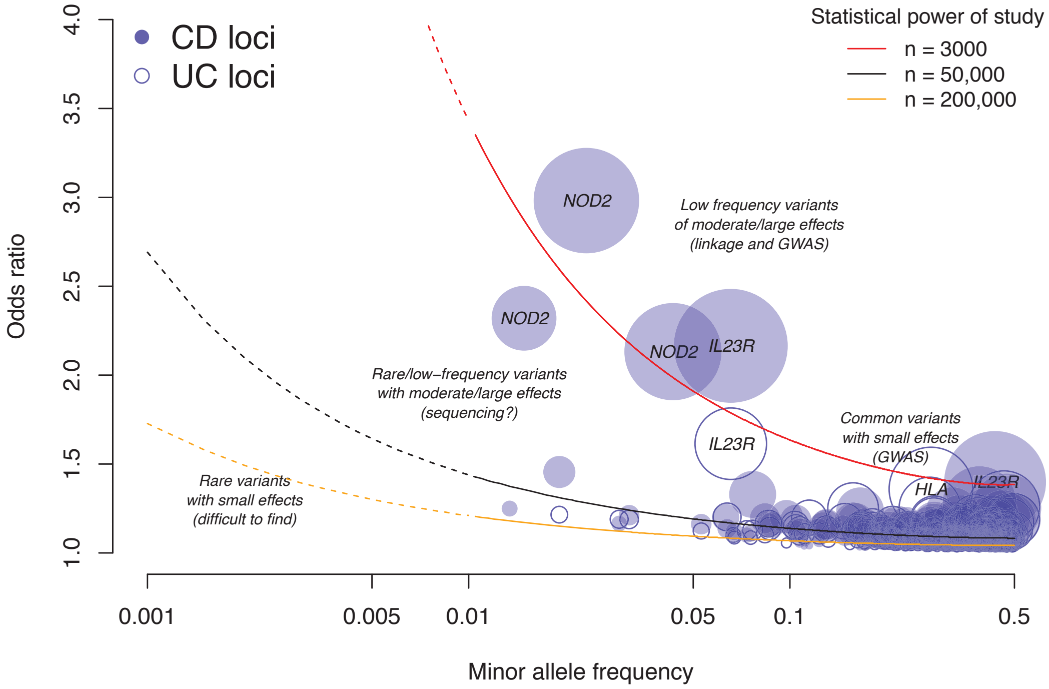

Figure 6.3. The genetic architecture of inflammatory bowel disease. Known CD and UC variants are plotted according to their minor allele frequencies and risk increasing ORs estimated from results in Chapter 4. ORs of risk-decreasing minor alleles were flipped for illustrative purposes. The size of the circles represents the amount variance in disease liability explained by that variant. The red, black and orange lines represent the minimum OR and allele frequency combination for a locus for which a GWAS with 3000, 50,000 or 200,000 individuals (with an equal number of cases and controls) respectively will have greater than 80% statistical power to detect association at $P < 5\times10^{-8}$. The dashed lines represent the allele frequency spectrum of variants that are typically poorly captured on GWAS microarrays (minor allele frequencies less than 1%).

The testing of rare variant associations will also throw up new statistical challenges. Firstly, the established genome-wide significance threshold of $P < 5\times10^{-8}$ is based on a 5% Bonferroni correction on the approximate number of independent regions tagged by common variants genome-wide (in European populations) (International HapMap *et al.,* 2007). Rare variants, on the other hand, are more numerous and less likely to be in high LD with other variants, such that a genome-wide survey of rare variants will involve many more independent tests than with common variants. It may be the case that a more stringent P-value threshold will be required to avoid too many false-positive reports. Secondly, while methods such as logistic regression are frequently used for common-variant associations, they may not be well-calibrated in rare-variant tests were minor allele counts are low, leading to false type-1 error rates. Ma *et al.* (2013) suggest a minor allele count cut-off of 400 (corresponding to a minor allele frequency of 1% in 20,000 individuals) for when standard logistic regression tests may need to be recalibrated. Thirdly, rare variants are more likely to be population specific, requiring more careful consideration of sample recruitment and study design. There is evidence that the effects of population stratification for rare variants are stronger than for common variants, and that existing methods such as PCA and linear mixed models may not be able to fully account for rare such stratification (Mathieson and McVean, 2012). Family-based association methods, which are robust to population stratification, may once again play an important role. Fourthly, different sequencing studies are likely to involve a range of sequencing technologies and methods, such that differences in coverage, read lengths, variant calling and genotype refinement methods will likely have direct effects on the properties of the variants reported. Methods that account for these differences, especially when cases and controls are sequenced separately, need to be developed. Finally, while I have discussed these challenges in the context of single-variant association tests, they also equally apply to the suite of rare-variant region-based tests. Additional challenges to these region-based methods include the choice of test, defining regions and which SNPs to test and difficulties in assigning causal variants (some of which I discussed in Chapter 1 section 1.4.3).

Despite these hurdles, there is a growing recognition among health policy makers about the importance of sequencing in medical research. In December 2012, the UK Government announced an initiative to sequence 100,000 whole genomes by the end of 2017. Patients will be recruited from NHS centres, and will consist of those with rare diseases or various cancers. The project is currently in its pilot phase, and it remains to be determined exactly how the samples will be sequenced, but will likely involve high-coverage sequencing of 40,000 patients. For rare diseases, the parents of patients will also be sequenced, and for each cancer patient, two genomes will be sequenced – one from the tumour and one from healthy tissue (Connor, 2014).

The immediate benefits of the UK 100K Project will be felt by patients and their families. For example, the sequencing of rare disease families will help in identifying highly penetrant de novo mutations, perhaps easing parents' concerns about having additional children. Identifying the somatic driver mutations in cancer can also inform the best course of treatment. For complex disease researchers, the project provides an invaluable resource to use with existing datasets (for instance, as population controls or imputation reference panels), as well as a testing ground for methods development. In the long term, the infrastructure, knowhow and experience gained through the project will provide a blueprint for future sequencing efforts. With the announcement of the Illumina HiSeq X Ten in January 2014, the raw cost of sequencing a genome at high-coverage (30X) today is around $1000. It is no stretch to imagine that this price will fall to a few hundred dollars within the decade, well below the price of many routine medical diagnostic tests, such that getting your genome sequenced (if you haven't already) will become part-and-parcel of a trip to the doctor.

What will it mean then, for complex disease research, when every patient with Crohn's disease, type 2 diabetes or schizophrenia in the country will have their genomes available? For locus discovery, the list of disease-associated variants will continue to grow. In a study of say, 300,000 Crohn's disease cases and a million controls, there will be greater than 80% power to detect variants with odds ratios greater than ~1.1 and a minor allele frequency of 1%. For variants with frequencies around 0.1%, odds ratios greater than ~1.4 will be detected. The total number of risk loci will likely be in the thousands, and, by this stage, insights into disease biology will primarily come from the molecular pathways and biological mechanisms that these risk loci cluster into and interact with rather than investigating the genes in isolation. By having entire families sequenced, it will also allow the identification of any *de novo* and rare highly penetrant variants in complex disease risk. In addition, such sample sizes along with medical records will also enable well-powered studies on disease subphenotypes such as clinical progression and drug response. Variants associated with disease progression may not necessarily also be associated with disease susceptibility (Lee *et al.,* 2013a), and such studies, along with integration with real-time monitoring of gut microbiota and cellular markers such as gene expression and epigenetic marks in disease-relevant cells, will pave the way for personalised treatments based on an individual's genetic makeup.

6.2.3 Genetic studies in non-European populations

As many as 96% of published GWAS up to 2011 were conducted in populations of European descent, yet these populations make up less than 15% of the world (Bustamante *et al.,* 2011). This disparity is primarily driven by resources – Western countries overwhelmingly spend more on scientific research, both in absolute terms and as a proportion of GDP, than do non-Western countries. It is then no surprise that the types of studies that rely on cutting edge technology (while costs are still at a premium) are first undertaken in these countries. Reassuringly, efforts such as the African Genomes Project, targeted funding efforts from research charities such as the Wellcome Trust, as well as the ever growing stream of home grown genetic studies emerging from researchers in Asian and Latin American countries are leading the charge in addressing this imbalance.

There is great scientific value in expanding complex trait genetics to the rest of the world. Firstly, as demonstrated for IBD in Chapter 4, much of the risk loci for complex disorders are likely to be shared across populations. This means that ascertaining samples from non-European populations is an effective way of boosting power to detect association. Of course, researchers will need to be aware of the potential for population stratification, though statistical methods that account for population stratification and potential heterogeneity between populations are now quite mature for common variant associations (Morris, 2011; Yang *et al.,* 2014a). Secondly, genetic differences between populations can inform biology. SNPs that are monomorphic in Europeans will go undetected in GWAS, yet finding associations at these variants in non-European populations will create new leads in understanding disease pathogenesis. When one considers the genetic diversity of African populations, this will almost certainly be the case. Moreover, variants that show large differences in effect sizes between populations point to potential gene-environment interactions, allowing for insights into the environmental factors that modify disease risk. Different population histories also create different patterns of LD. These patterns will be instrumental in fine-mapping efforts to localise causal variants at associated loci common across populations. Finally, aside from the scientific reasons listed above, there are clear humanitarian arguments for expanding genetic studies to non-Western countries and to study the diseases that most burden them. Those most in need must not be the last to benefit from genetic research (Bustamante *et al.,* 2011).

6.2.4 Genetic prediction

In addition to gaining a better understanding of disease biology, genetic information can also potentially be used for disease risk prediction. Prediction methods for complex diseases typically involve assigning a risk score to an individual based on their genotypes and previously estimated effect sizes (for instance, ORs from GWAS) across risk alleles. Risk alleles can be assigned not only based on known associations, but also include nominally associated variants. Prediction accuracy can be evaluated by methods such as the receiver operating characteristic curve (ROC), which estimates the true and false positive rates of the predictor at various risk score cut-offs (Lasko *et al.,* 2005). The area under the ROC (AUC) is the probability that for a randomly selected pair of diseased and healthy individuals, the diseased individual will have a higher risk score. An AUC of 0.5 means that the prediction method is no better than chance, while a value of 1 means that the method perfectly discriminates between diseased and healthy individuals.

For complex autoimmune diseases, genetic risk prediction is still in its infancy and does not currently offer much in terms of clinical utility. Estimates of

AUC using just family history of disease, genetic risk loci or the two together in Crohn's disease range from 0.56 to 0.74 (Kang *et al.,* 2011; Ruderfer *et al.,* 2010). Including risk factors such as smoking and age into the risk model may improve the AUC. Nevertheless, given its high heritability, the theoretical maximum possible AUC assuming that all Crohn's disease risk loci have been identified and effect sizes are accurately measured is estimated to lie between 0.96-0.98 (Jostins and Barrett, 2011; Wray *et al.,* 2010). However, while this figure seems high, the utility of genetic prediction is limited given the low prevalence of Crohn's and other immune-mediated diseases. Even assuming a generous disease prevalence estimate of 1% and AUC of 0.98, less than 12% of individuals who test positive (using a sensitivity cut-off of 0.93) will develop disease (Jostins and Barrett, 2011). Increasing the threshold will increase the proportion of positively identified individuals but also exclude a higher number of cases from being identified. While never providing any guarantees, the use of genetic prediction in complex diseases may ultimately at best aid in disease diagnosis, and at worst create greater awareness among those most highly at risk for disease.

6.3 From causal variants to treatment outcomes

In Chapter 1, I discussed potential approaches and challenges involved in narrowing down a risk locus into a single causal variant. Assuming now that a set of causal candidates has been identified, what is required to confirm causality? The direct modelling of these variants in cell lines and model organisms are likely to play an important role in answering this question, and emerging technologies such as DNA editing through CRISPR/Cas and engineering induced pluripotent stem cells (iPSCs) are growing in popularity (Cong *et al.,* 2013; Mali *et al.,* 2013; Robinton and Daley, 2012). The CRISPR/Cas system involves guiding a Cas-cleavage enzyme to a specific site of the genome, which is then imprecisely cleaved and repaired, allowing for specific mutations to be introduced. *In vitro* modelling of these mutations in disease relevant cells types (e.g. those generated from iPSCs) allows for the direct investigation of how these mutations affect cellular phenotypes such as gene expression and responses to infection;

generating hypotheses about how these genetic variants lead to disease susceptibility. Knocking down the relevant genes identified in model organisms will further enable understanding of how these genes affect the organism as a whole.

At this point, it is worth discussing about the level of proof required before causality can be confidently assigned to a genetic variant. From a genetic association standpoint, defining causality is straightforward, though identifying it is difficult. A causal variant is one that can explain a statistical association signal on its own, irrespective of its correlation with other variants. Hence an associated tag SNP cannot be called causal. From a disease risk standpoint, however, causality is more nuanced. Given the typical small effect sizes of associated variants, a causal variant is neither necessary nor sufficient to cause disease (Visscher *et al.,* 2012). CRISPR/Cas, iPSCs and gene knockouts might reveal a disease relevant phenotype, but this also does not prove that the phenotype affects disease risk in the population. It may be the case that we will never have the ability to definitively prove that the observed biological effect of a statistically causal variant is also causal in the disease risk sense.

Defining causality may end up being a moot point if the relevant genes that are identified and biological knowledge gained lead to better treatment outcomes. Identifying a gene target and the creation of a therapeutic molecule is difficult. Over 90% of compounds that enter clinical trials fail to gain approval, reflecting the limited predictive value of preclinical disease models and a lack of understanding of the long-term consequences of perturbing specific molecules (Plenge *et al.,* 2013). While GWAS have provided valuable insights into disease biology, little of this has yet translated into more effective therapeutics. Part of this is of course due to time – moving from a gene target through clinical trials to a final approved drug can take well over a decade. Nevertheless, it is hoped that knowledge of the genes that underlie disease risk will lead to more effective treatment outcomes.

There is a strong historical precedence for the use of human genetics in drug development. Before the large-scale identification of susceptibility genes, epidemiological observations were often the catalyst for identifying potential therapies. Genetic variation in the human population meant that many individuals carried alleles that mimic the effects of potential therapies. For instance, the development of statins to lower LDL cholesterol levels and treat heart disease was based on observations in families with rare hypercholesterolemia who carried mutations in the *LDLR* gene. Members of these families both had higher levels of cholesterol and higher prevalence of heart disease. Importantly, the number of mutations appeared to affect cholesterol levels and risk of heart disease in a dose-dependant manner. It was also known that the HMG-CoA reductase plays an important role in the production of cholesterol in the liver, and natural products that inhibit this enzyme (e.g. compactin and lovastatin) lowered LDL cholesterol levels in animal models (Plenge *et al.,* 2013). Later clinical trials in humans demonstrated the efficacy and safety of statins, and ultimately showed their effectiveness at reducing heart disease risk in individuals with high cholesterol.

The role of human genetics in drug development is also supported retrospectively by drugs that were developed without the use of human genetics, but whose molecular targets have since been supported by their associations with disease. In the statins example, variants in the *HMGCR* gene (which encodes the HMG-CoA enzyme) were found to be associated with LDL cholesterol by GWAS (Kathiresan *et al.,* 2008). Notably, the effect size of the association bears little relationship to its clinical relevance. The HMGCR signal has an effect on LDL cholesterol levels of ~2.5 mg/dl per allele (Teslovich *et al.,* 2010), or, to put another way, approximately one-tenth of a unit of standard deviation – a tiny effect. Yet statin drugs can reduce LDL levels by around 40 mg/dl (Cholestrol Treatment Triallists' Collaborators, 2005). Other retrospective examples include the targeting of *CTLA-4* by abatacept for rheumatoid arthritis (Genovese *et al.,* 2005; Gregersen *et al.,* 2009), *IL12B* by ustekinumab for Crohn's disease (Mannon *et al.,* 2004; Parkes *et al.,* 2007) and *PPARG* by thiazolidinediones for type 2 diabetes (Spiegelman, 1998; Zeggini *et al.,* 2007).

The identification of shared risk loci across different diseases also enables repurposing of existing drugs. By looking at the overlap between GWAS risk loci and current drugs in development, Sanseau *et al.* (2012) identified over 100 targets that were associated with a disease other than the one the drug was being developed for. For instance, *TNFSF11* is currently inhibited by denosumab for treatment of osteoporosis in postmenopausal women. Variants in *TNFSF11* are also associated with Crohn's disease (Franke *et al.,* 2010), and it is tempting to suggest that this drug may be repurposed (Sanseau *et al.,* 2012). The use of existing approved drugs also avoids the need for lengthy safety trials, meaning that treatments can be marketed in a much shorter time frame.

Using GWAS risk loci to guide the development of novel drugs will be a much bigger challenge. Plenge *et al.* (2013) list nine criteria for prioritising risk loci before drug discovery should be considered:

1. The gene harbours a causal variant that is unequivocally associated with a medical trait of interest
2. The biological function of the causal gene and causal variant are known
3. The gene harbours multiple causal variants of known biological function, thereby enabling the generation of genotype–phenotype dose–response curves
4. The gene harbours a loss-of-function allele that protects against disease, or a gain-of-function allele that increases the risk of disease
5. The genetic trait is related to the clinical indication targeted for treatment
6. The causal variant is associated with an intermediate phenotype that can be used as a biomarker
7. The gene target is druggable
8. The causal variant is not associated with other adverse event phenotypes
9. Corroborating biological data support genetic findings

Going through this list, it's clear that, with the exception of point 1, the majority of GWAS risk loci do not yet satisfy any of these criteria. The rationale for many of these points (e.g. the need for loss-of-function or gain-of-function alleles) is that it is simply easier to develop drugs that inhibit certain type of targets with the current knowledge of assays. This will hopefully change in the future as technology and assays improve. For instance, kinases were once thought to be undruggable, though this is now changing with the development of kinase-inhibitors (Gashaw *et al.,* 2011). Moreover, some GWAS risk loci themselves may not necessarily be the most actionable drug target, but rather can inform molecular pathways that are relevant to disease. For instance, if a risk locus is a ligand, knowing the corresponding receptor (for which drugs are well-suited to exert their effects on) will offer additional potential targets. Ultimately however, understanding the disease-relevant biological functions of risk loci will always remain the first step on the road to drug discovery.

6.4 Concluding remarks

It is almost certain that within the coming decades, low-cost whole genome-sequencing will become routine, and it is not too much of a stretch to imagine locus discovery projects involving hundreds of thousands, perhaps even millions, of whole-genome sequenced cases and controls. While the theoretical framework of association studies as outlined in Risch and Merikengas (1996) are unlikely to change, these types of studies will also throw up new methodological challenges that will need to be overcome. Along with genome-sequencing, large scale functional genomic studies will ever expand to include greater coverage of cell types and disease states, and methods to integrate these data sources will play an important role in understanding biology.

It needs to be emphasised that locus discovery is not an end in itself. Challenges remain in taking what we’ve learned from genetic studies to build more complete models of disease pathogenesis and ultimately translating these into better patient outcomes.